بیگ‌لو، اعظم؛ داورپناه، محمدرضا (۱۳۹۵). محاسبه بار اطلاعاتی متون علمی فارسی براساس شاخص آنتروپی نظریه اطلاعات. *پژوهش‌نامه کتابداری واطلاع‌رسانی*، ۶ (۱)، ۱۰۸– ۸۸.

# محاسبه بار اطلاعاتی متون علمی فارسی براساس شاخص آنتروپی نظریه اطلاعات [۱]

اعظم بیگ‌لو [۲]، دکتر محمدرضا داورپناه [۳]

تاریخ دریافت: ۹۲/۶/۳۰ تاریخ پذیرش: ۹۲/۹/۲

**چکیده**

***هدف:*** *هدف عمده این پژوهش، محاسبه میزان بار اطلاعاتی واژه‌های متون علمی فارسی و بررسی رابطه برخی ویژگی‌های واژه و بار اطلاعاتی آن برمبنای مقیاس آنتروپی شانون است.*

***روش:*** *پژوهش حاضر با روش تحلیل محتوا و در جامعه آماری شامل ۷۵۲ مقاله برگرفته از فهرست مجلات علمی پژوهشی در سال ۱۳۸۸ صورت پذیرفت. نمونه پژوهش شامل ۳۲۰ مقاله بود که با توجه به گستردگی آن در هر حوزه تنها ۱۰ درصد از مقالات به صورت تصادفی انتخاب و مورد بررسی قرار گرفت.*

***یافته‌ها:*** *پژوهش حاضر نشان داد، بار اطلاعاتی واژه با احتمال رخداد آن رابطه‌ای معکوس دارد. با افزایش تعداد حالات ممکن میزان پیش‌بینی‌پذیری و آنتروپی واژه افزایش یافته و اطلاعات کمتری منتقل می‌نماید. طول واژه نیز رابطه‌ای مستقیم با بار اطلاعاتی آن دارد. حوزه‌های مختلف علمی در میزان اطلاعاتی که انتقال می‌دهند یکسان نیستند و حوزه علوم انسانی بیشترین میزان آنتروپی و کمترین میزان اطلاعات را نسبت به سایر حوزه‌ها داراست.*

***کلیدواژه‌ها:*** *نظریه اطلاعات، آنتروپی، بار اطلاعاتی واژه، متون علمی فارسی.*

---

۱. برگرفته از پایان نامه کارشناسی ارشد

۲. کارشناس ارشد علم اطلاعات و دانش‌شناسی دانشگاه فردوسی مشهد، beiglooazam@gmail.com

۳. استاد گروه علم اطلاعات و دانش‌شناسی دانشگاه فردوسی مشهد، davarpanah@um.ac.ir

## مقدمه

پیشرفت روزافزون نظام‌های ذخیره و بازیابی اطلاعات و نیاز به بهینه نمودن بازیابی اطلاعات از مدارک متنی باعث شده است تا در سال‌های اخیر توجه گسترده‌تری معطوف به فنون و رویکردها در زمینه نظام‌های نمایه‌سازی خودکار شود. پیامد استفاده از نظام‌های نمایه‌سازی مبتنی بر زبان طبیعی انعکاس انواع کلمات در نمایه است. از آنجا که تمام واژگان در متن ارزش و بار اطلاعاتی یکسانی ندارند استفاده از روش‌هایی که کلمات مهم را از کلمات بی‌اهمیت تشخیص دهد همیشه در این حوزه مورد توجه بوده است. گروهی از واژگان زبان طبیعی (مانند حروف تعریف، حروف ربط، حروف اضافه، و برخی از افعال) سهم معنایی یا دستور زبانی بسیار پایینی دارند. به عبارت دیگر، حشو ویژگی بارز متون زبان طبیعی است که به منظور جلوگیری از اختلال در درک پیام متن به کار می‌رود. از سوی دیگر، بسامد واژه به تنهایی، به منظور اختصاص کلیدواژه‌های موضوعی مدارک چندان قابل اعتماد به نظر نمی‌رسد ( داورپناه و بلندیان، ۱۳۸۶). علاوه براین، اگر فرایند استخراج کلیدواژه بدون توجه به بار اطلاعاتی و وزن معنایی کلمه انجام پذیرد، علاوه بر حجیم شدن پایگاه واژگان نمایه، ریزش کاذب و بازیابی منابع نامرتبط نیز دور از انتظار نخواهد بود، چرا که واژگان اطلاعی بار اطلاعاتی یکسان ندارند. درواقع هر کلمه یا ترکیب به یک میزان اطلاع‌دهنده[1] نیستند، یعنی بسیاری از ترکیبات و کلمات نباید در نظام‌های نمایه‌سازی به عنوان واژه‌نما انتخاب شوند.

بدیهی است که همه واژگان در زبان نوشتاری مقادیر یکسانی از اطلاعات را انتقال نمی‌دهند. لوهن[2] (۱۹۵۷) نخستین کسی بود که بیان کرد واژگانی که با فراوانی بسیاری رخ می‌دهند سهم معنایی عمده‌ای در یک متن ندارند. علاوه بر این، این واژگان بخش بزرگی از متن، سهمی حدود ۲۰ تا ۳۰ درصد از نشانه‌ها در متن را، به خود اختصاص می‌دهند. از سوی دیگر، تقریباً نیمی از واژگان تنها یک‌بار در یک پیکره متنی[3] رخ می‌دهند؛ درحالی‌که بیشتر واژگان حدود ده بار یا کمتر در متن ظاهر می‌شوند ( Manning & Schutze,1999). حتی پس از استخراج سیاهه بازدارنده، واژگان بسیار زیادی باقی خواهد ماند که همگی به عنوان واژه‌نما قابل توجه نیستند. بنابراین لازم است با استفاده از سایر روش‌ها، سودمندی واژگان باقی‌مانده تعیین شود. مطالعات فراوانی در زبان انگلیسی به رتبه‌بندی اطلاع‌دهی[4] واژه پرداختند. منظور از اطلاع‌دهی واژه این است که آن واژه به چه میزان بیانگر ایده‌های کلیدی در مجموعه‌ای از مدارک است

1. Informative
2. Luhn
3. Corpus
4. Informativeness

(Tomokiyo & Hurst, 2003). به عبارت دیگر، اطلاع‌دهی واژه نشان‌دهنده درجه و اندازه‌ای است که عبارت کلیدی معرف و نمایانگر مدرک در دست بررسی است، و با میزان اطلاعاتی که به کاربر منتقل می‌نماید همبسته است.

تاکنون روش‌های مختلفی برای اندازه‌گیری بار اطلاعاتی واژه به کار گرفته شده است. از جمله این روش‌ها، استفاده از کمیت آنتروپی نظریه ریاضی اطلاعات[1] شانون[2] می‌باشد. این نظریه شاخه‌ای از نظریه آماری علوم ارتباطی است و شیوه‌ای کمی برای اندازه‌گیری محتوای اطلاعاتی پیام‌ها به دست می‌دهد. نظریه اطلاعات به طور عمده ناظر بر مسئله تعیین حداکثر ظرفیت یک کانال یا یک مجرا برای انتقال پیام‌هاست و هدف اصلی شانون دست یافتن به شیوه‌ای بود که کارایی کانال ارتباطی را به حداکثر برساند (حری، ۱۳۸۱، ص ۲۳۲). نظریه اطلاعات با ارائه شاخصی به نام آنتروپی به اندازه‌گیری اطلاعات یک متغیر تصادفی می‌پردازد. این متغیر تصادفی می‌تواند واحدی از یک متن (حرف، کلمه، جمله، و ...) باشد. با استفاده از آنتروپی می‌توان میزان محتوای اطلاعاتی[3] یک واژه را به عنوان متغیری تصادفی اندازه گرفت و کلمات با آنتروپی بالا را نادیده گرفت؛ به همین ترتیب می‌توان سیاهه‌ای از واژگان غیرمجاز ساخت. مطالعات نشان داده است که استفاده از شاخص آنتروپی نسبت به سایر معیارهای نحوی برای تشخیص و شناسایی واژگان کارکردی و ساخت سیاهه واژگان غیرمجاز مفیدتر است (Melamed, 1997). امروزه استفاده از شاخص آنتروپی جایگزین روش‌های نمایه‌سازی واژگان براساس آمارهای فراوانی واژه شده است. روش‌های مبتنی بر فراوانی سطحی هستند و مفهوم اصلی متن را منعکس نمی‌کنند، در حالی‌که مدل‌هایی مبتنی بر آنتروپی دقت بالایی را گزارش کردند (Fragos, Maistros & Skourlas, 2005; Lin & Yu-Shu, 2005). براساس آنچه گفته شد، مسئله اساسی این پژوهش آن است که برمبنای آنتروپی نظریه اطلاعات بار اطلاعاتی کلمات در متون علمی فارسی چگونه است؟

## هدف‌ها و ضرورت پژوهش

هدف اصلی این پژوهش محاسبه میزان آنتروپی متون علمی و تخصصی زبان فارسی است. سایر اهداف پژوهش حاضر شامل محاسبه بار اطلاعاتی واژگان متن و شناسایی واژگان کم بار اطلاعاتی (واژگان غیرمجاز)، بررسی رابطه بار اطلاعاتی یک واژه با احتمال رخداد، تعداد حالات ممکن، طول واژه

1. Mathematical Information Theory
2. CludeShanon
3. Information Content

و زمینه (حوزه) آن است. با توجه به لزوم بهبود نظام‌های بازیابی اطلاعات و نمایه‌سازی مدارک و نیز برخی ویژگی‌های خاص خط و زبان فارسی و نظر به آنکه در زبان فارسی کمتر به موضوع وزن‌دهی واژگان پرداخته شده است، لزوم پژوهشی در راستای تعیین میزان اهمیت واژه، می‌تواند در حوزه بازیابی اطلاعات و نظام‌های نمایه‌سازی راهگشا باشد.

## فرضیه‌های پژوهش

۱. میزان بار اطلاعاتی واژه با احتمال رخداد آن رابطه‌ای معکوس دارد.
۲. هرچه میزان آنتروپی متن بیشتر باشد میزان حضور اطلاعات در متن کمتر است.
۳. هرچه تعداد حالات ممکن یک واژه کمتر باشد بار اطلاعاتی آن واژه بیشتر است.
۴. بین طول کلمه و بار اطلاعاتی آن رابطه وجود دارد.
۵. میزان بار اطلاعاتی متون در حوزه‌های مختلف علمی متفاوت می‌باشد.

## پیشینه پژوهش

محاسبه بار اطلاعاتی واژه براساس آنتروپی در پژوهش‌های بسیار و با اهداف مختلفی مورد توجه قرار گرفته است. کاربرد آنتروپی در تحلیل و مطالعه متون سابقه‌ای طولانی دارد. در ادامه به پژوهش‌هایی که میزان اطلاعات واژه را براساس آنتروپی سنجیده‌اند اشاره می‌کنیم.

یکی از هدف‌های تعیین بار اطلاعاتی واژه، شناسایی واژه‌های عمومی و اختصاصی یک متن است؛ میزان اطلاعات و بنابراین، آنتروپی واژه‌های عمومی و تخصصی در متون یکسان نیست و این موضوع در پژوهش‌های بسیاری مورد بررسی قرار گرفته است. کارابالو و چارنیاک[۱] (۱۹۹۹) از طریق محاسبه میزان آنتروپی کلمات سطح اختصاصی بودن[۲] آن‌ها را مورد اندازه‌گیری قرار دادند. نتایج نشان داد اسم‌هایی با آنتروپی و فراوانی بیشتر عمومی‌تر هستند و می‌توان گفت رابطه معکوسی بین فراوانی و محتوای معنایی یک کلمه وجود دارد. پژوهش دیگری با هدف شناسایی واژگان اختصاصی، پژوهش ریو و چوی[۳] (۲۰۰۴) است که با بررسی اصطلاحات اصطلاحنامه مش[۴] دریافتند آنتروپی با دقتی بسیار بالا (۸۲ درصد) به شناسایی اصطلاحات تخصصی می‌پردازد. نمیروفسکی و دوبراینین[۵] (۲۰۰۸) به بررسی اهمیت

1. Caraballo & Charniak
2. Specificity
3. Ryu & Choi
4. MeSH
5. Nemirovsky & Dobrynin

واژه در بافت و زمینه‌ای که در آن رخ می‌دهد، پرداختند. نتایج نشان داد واژه‌ای مهم است که در یک حوزه معنای تخصصی داشته باشد و در تعداد مدارک اندکی رخ دهد. به عبارت دیگر واژه تخصصی در متن آنتروپی پایینی دارد. درنهایت، می‌توان به پژوهش کیریو[۱] (۲۰۰۹) اشاره کرد که او نیز مشابه پژوهش‌های قبل، اختصاصی بودن، بار اطلاعاتی و محتوای اطلاعاتی واژه بر اساس آنتروپی را مورد بررسی قرار داد. او نشان داد به‌کارگیری مدل آنتروپی نتایج بسیار بهتری نسبت به روش‌های سنتی وزن دهی کلمات در پی دارد.

گروه دیگری از مطالعات به بررسی بار اطلاعاتی واژه پرداختند. از جمله این پژوهش‌ها پژوهش مونتمورو و زانت[۲] (۲۰۰۱) است که بیش از سایر پژوهش‌ها به مطالعه حاضر نزدیک و با آن همسو می‌باشد. پژوهشگران در مطالعه خود به تحلیل آماری کلمات در مجموعه متون ادبی انگلیسی پرداختند. نتایج نشان داد آنتروپی با افزایش تکرار کلمه افزایش می‌یابد، به عبارتی، کلمات کمیاب آنتروپی بسیار پایینی دارند. گنزل و چارنیاک[۳] (۲۰۰۲) با بررسی متون وال استریت ژورنال[٤] از سال ۱۹۸۷ تا ۱۹۸۹ به این نتیجه که «جملات در میزان اطلاعاتی که انتقال می‌دهند متفاوت هستند»، دست یافت.

تعیین واژگان بازدارنده نیز یکی از هدف‌های محاسبه بار اطلاعاتی واژه به حساب می‌آید. زو و همکارانش[۵] (۲۰۰۶) مدلی مبتنی بر نظریه اطلاعات به منظور ساخت سیاهه واژگان بازدارنده برای متون چینی ارائه کردند. از دیدگاه نظریه اطلاعات واژگان بازدارنده واژگانی هستند که اطلاعات کمی را انتقال می‌دهند. ارزش اطلاعاتی واژه $wj$ توسط آنتروپی اندازه‌گیری می‌شود. احتمال $Pi,j$ فراوانی آن در مدرک $Di$ تقسیم بر تعداد کل واژگان در مدرک $Di$ است. بنابراین، ارزش آنتروپی $(H)$ برای واژه $wj$ به صورت زیر محاسبه گردید:

$$H(w_j) = \sum_{N}^{i=1} P_{i,j} \times \log\left(\frac{1}{P_{i,j}}\right)$$

پژوهش‌هایی که به طور مستقیم به اندازه‌گیری «آنتروپی واژه» پرداخته‌اند در زبان فارسی چندان پرشمار نیستند.آنتروپی حروف فارسی در مطالعه میرزایی (۱۳۸۵) مورد توجه قرار گرفت. هدف این

1. Kireyev
2. Montemurro & Zanette
3. Genzel & charniak
4. The Wall Street Journal
5. Zou et al.

مطالعه بررسی حشو در زبان فارسی با رویکرد نظریه اطلاعات بود. در نهایت این گونه نتیجه‌گیری شد که آنتروپی زبان فارسی روی حرف اول کلمه ۵ بیت است و زوائد آن حدود ۸۰ درصد می‌باشد. بنابراین اختیار ما برای سخن گفتن ۲۰ درصد و احاطه زبان فارسی بر گفتار ما حدود ۸۰ درصد است.

مطالعه بر نرخ آنتروپی زبان فارسی در پژوهش هاشمی و ساوجی (۱۳۸۶) نیز صورت پذیرفت. با استفاده از نتایج به دست آمده از مدل‌سازی متن فارسی و انگلیسی مشخص شد نرخ آنتروپی متن فارسی بالاتر از متن انگلیسی و در نتیجه قابلیت فشرده پذیری متن فارسی کمتر از متن انگلیسی است. در پژوهش‌های ذکر شده اشاره‌ای به بار اطلاعاتی نشده است، در حالی که، حجازی، علی و عابد[۱] (۱۹۸۷) در مطالعه خود به اندازه‌گیری بار اطلاعاتی هر حرف و ترکیباتی از چند حرف پرداخته و سپس این میزان اطلاعات را با زبان انگلیسی مقایسه نمودند. در نهایت چنین استنتاج شد که وقتی ارزش n (طول کلمه) افزایش می‌یابد، آنتروپی کاهش می‌یابد.

با نگاهی به مطالعات مربوط به آنتروپی واژه و بار اطلاعاتی آن می‌توان دریافت که هدف عمده این پژوهش‌ها شناسایی واژه‌های عمومی و اختصاصی، واژگان بازدارنده (غیرمجاز) و تعیین رابطه بار اطلاعاتی واژه با ویژگی‌های آن از جمله طول واژه، فراوانی و ... است؛ البته قابل ذکر است که این اهداف جملگی، بر بازیابی اطلاعات بهینه تأکید دارند. البته بررسی بار اطلاعاتی واژه، صرف نظر از هدف آن، موضوعی بود که در پژوهش‌های انجام شده در ایران به چشم نخورد و به نظر می‌رسد تنها، آنتروپی به عنوان کلیتی در ارتباط با زبان مورد توجه قرار گرفته است، درحالی‌که بازیابی اطلاعات و نمایه‌سازی کارآمد، به انجام مطالعات گسترده‌تر و دقیق‌تری بر بار اطلاعاتی واژه متکی است.

## روش‌شناسی پژوهش

این پژوهش با روش تحلیل محتوا انجام پذیرفت. وبر[۲] (۱۹۹۰) تحلیل محتوا را روش تحقیقی قاعده‌مند به منظور تحلیل اطلاعات متنی طبق یک روش استاندارد و تعریف شده معرفی می‌نماید که به محقق اجازه می‌دهد تا براساس این اطلاعات، نتیجه‌گیری کند. رویکردهای کمی تحلیل متن با استفاده از روش‌های آماری و ریاضی صورت می‌پذیرد و از آنجا که هدف این پژوهش استفاده از شاخص آنتروپی به منظور تحلیل متون فارسی است، روش تحلیل محتوا روشی مناسب می‌باشد.

1. Hegazi, Ali & Abed
2. Weber

جامعه آماری این پژوهش مقالات مندرج در آخرین شماره منتشر شده در مجلات علمی- پژوهشی حوزه‌های ادبیات و علوم انسانی، علوم پایه، فنی و مهندسی، و کشاورزی است. این مجلات که از فهرست نشریات مورد تائید وزارت علوم، تحقیقات و فناوری در سال ۱۳۸۸ استخراج شد، شامل ۲۶۱ عنوان مجله در چهار حوزه است. تعداد مقالات منتشر شده در آخرین شماره این مجلات نیز حدود ۲۶۵۰ مقاله بود که از این تعداد مقاله تنها عناوینی مورد توجه قرار گرفت که به صورت الکترونیکی در دسترس بودند. به این صورت تعداد ۸۳ مجله و ۷۵۲ مقاله در حوزه علوم انسانی، ۲۲ مجله و ۱۷۰ مقاله در حوزه فنی و مهندسی، ۳۸ مجله و ۴۹۵ مقاله در حوزه کشاورزی و منابع طبیعی و ۹ مجله و ۹۸ مقاله در حوزه علوم پایه به عنوان جامعه آماری پژوهش مورد توجه قرار گرفت، سپس با استفاده از فرمول زیر تعداد مقالاتی که باید در هر حوزه مورد بررسی قرار گیرند محاسبه شد:

$$\text{تعداد مقالات هر حوزه} = \frac{\text{حجم جامعه آن حوزه}}{\text{حجم کل جامعه}} \times \text{حجم نمونه}$$

نتایج حاصل از این معادله تعیین کرد که تعداد ۱۵۴ مقاله در حوزه علوم انسانی، ۳۵ مقاله در حوزه فنی و مهندسی، ۱۰۱ مقاله در حوزه کشاورزی و منابع طبیعی، و ۲۰ مقاله در حوزه علوم پایه باید مورد بررسی قرار گیرند. با توجه به لزوم مقایسه متون در حوزه‌های مختلف و به‌کارگیری روش‌های آماری، حجم نمونه در حوزه علوم پایه نیز به ۳۰ عنوان افزایش یافت. شناسایی این مقالات از میان کل تعداد مقالات در هر حوزه به صورت تصادفی ساده انجام پذیرفت. قابل ذکر است در فرضیاتی که در سطح مقاله باید آزمون می‌شد تنها از ده درصد مقالات یعنی ۳۲ مقاله استفاده شد و این تعداد نیز به صورت تصادفی ساده انتخاب شدند.

## گردآوری داده‌ها

به منظور تحلیل متون مورد مطالعه، ابتدا متن الکترونیکی متون به قالب Word تبدیل و در این محیط نرم‌افزاری تحلیل و پردازش واژگان امکان‌پذیر شد.

پس از تهیه متون الکترونیکی، واژگان تفکیک گردید. تفکیک واژگان در دو مرحله صورت پذیرفت:

۱. اول واژگان هر متن از نظر شیوه نگارش مورد ویرایش قرار گرفت. در این مرحله براساس آئین نگارش زبان فارسی، واژگان ساده و مرکب تشخیص داده شده و متمایز شدند. معیارهای مورد

استفاده در این مرحله معیارهای مورد استفاده در پژوهش‌های پیشین (داورپناه و بلندیان، ۱۳۸۶؛ سنجی و داورپناه، ۱۳۸۸) و قواعد نگارش متون فارسی ( وحیدیان کامیار، ۱۳۷۹) است.

۲. در مرحله دوم، متن مورد نظر را در محیط نرم‌افزاریWord و با استفاده از گزینه Table، به جدول تبدیل شد.

پس از تفکیک واژگان مقالات، برای هر مقاله یک جدول با چند ستون طراحی گردید. ستون اول مربوط به فراوانی واژه، ستون دوم طول واژه، ستون سوم حالات ممکن، ستون چهارم احتمال رخداد واژه و ستون پنجم آنتروپی بود. شمارش فراوانی واژگان به منظور محاسبه احتمال رخداد آن لازم است. تعداد کل واژه‌های هر مقاله و طول واژه در نوار وضعیت با استفاده از فعال نمودن قابلیت Word Count قابل مشاهده است. حالات ممکن یک واژه، شمارش یک واژه با کلیه کلمه‌هایهم‌خانواده و هم‌ریشه بود. آنتروپی واژه نیز به منظور برآورد باراطلاعاتی واژگان مورد توجه قرار گرفت.

## یافته‌های پژوهش

با استفاده از داده‌های گردآوری شده فرضیات پژوهش مورد بررسی قرار گرفت که توضیحات آن ذیل فرضیات ارائه می‌شود:

**فرضیه شماره ۱.** میزان بار اطلاعاتی واژه با احتمال رخداد آن رابطه‌ای معکوس دارد.

هدف از طرح این فرضیه بررسی رابطه بین احتمال رخداد یک واژه در یک متن و میزان اطلاعات آن است، به این صورت که واژه‌ای با احتمال وقوع مشخص نسبت به سایر واژه‌ها، بار اطلاعاتی کمتر یا بیشتری خواهد داشت. متغیرهای مورد بررسی در این فرضیه شامل احتمال رخداد ($P_I$) و بار اطلاعاتی است. به منظور بررسی بار اطلاعاتی واژگان در این پژوهش از مفهوم کمیت آنتروپی استفاده گردید. از آنجا که رابطه آنتروپی و اطلاعات یک رابطه معکوس می‌باشد آنتروپی بالای هر واژه نشانگر بار اطلاعاتی اندک آن است. پس از ویرایش متون، شمارش واژه‌های هر مقاله و محاسبه فراوانی ($f_i$)، احتمال رخداد تک تک واژه‌ها محاسبه گردید به این صورت که فراوانی و تعداد تکرار یک واژه در یک مقاله به تعداد کل واژگان آن مقاله ($N$) تقسیم شد:

$$Pi = \frac{fi}{N}$$

داده‌های مربوط به هر مقاله شامل فراوانی واژه، وارد برنامه اکسل شد و سپس با استفاده از نرم‌افزار آماری SPSS احتمال رخداد، لگاریتم احتمال رخداد و حاصل ضرب این دو کمیت محاسبه و طبق معادله

زیر آنتروپی واژه به دست آمد:

$$H = -P(i)LogP(i)$$

در مرحله آزمون فرض به منظور بررسی این رابطه از آزمون همبستگی پیرسون استفاده شد. ضریب همبستگی نشانه وجود ارتباط بین دو متغیر است. قابل ذکر است که هرچه ارتباط دو متغیر شدید باشد مقدار ضریب به ۱+ و ۱- نزدیک‌تر خواهد بود و با کاهش ارتباط بین دو متغیر مقدار ضریب به صفر نزدیک می‌شود. همچنین اگر دو متغیر مستقل باشند مقدار ضریب همبستگی برابر صفر خواهد بود.

همان‌طور که در جدول شماره ۱ مشاهده می‌شود، مقدار P در همه مقالات کمتر از ۰/۰۵ است بنابراین فرض آماری رد و فرض پژوهش پذیرفته می‌شود، به این معنی که میزان بار اطلاعاتی واژه متن با احتمال رخداد آن رابطه دارد. از طرفی مقدار ضریب همبستگی پیرسون نزدیک به ۱ و علامت آن مثبت است که نشان‌دهنده همبستگی بالا میان دو متغیر احتمال رخداد و آنتروپی می‌باشد.

**جدول شماره ۱. بررسی ضریب همبستگی آنتروپی واژه با احتمال رخداد**

| ردیف | کد مقاله | ضریب همبستگی | Sig. (2-tailed) | تعداد واژه |
|---|---|---|---|---|
| ۱ | ۳ | ۰/۹۸۴ | ۰ | ۹۳۹ |
| ۲ | ۴ | ۰/۹۷۱ | ۰ | ۱۱۹۲ |
| ۳ | ۶ | ۰/۹۷۲ | ۰ | ۸۷۳ |
| ۴ | ۷ | ۰/۹۷۴ | ۰ | ۸۵۵ |
| ۵ | ۱۰ | ۰/۹۷۷ | ۰ | ۷۰۰ |
| ۶ | ۱۷ | ۰/۹۸۲ | ۰ | ۵۰۰ |
| ۷ | ۲۰ | ۰/۹۷۹ | ۰ | ۸۵۸ |
| ۸ | ۲۱ | ۰/۹۸۲ | ۰ | ۱۵۶۶ |
| ۹ | ۲۳ | ۰/۹۷۴ | ۰ | ۵۴۹ |
| ۱۰ | ۲۷ | ۰/۹۷۳ | ۰ | ۶۰۹ |
| ۱۱ | ۲۸ | ۰/۹۳۸ | ۰ | ۶۸۶ |
| ۱۲ | ۳۰ | ۰/۹۷۸ | ۰ | ۱۲۲۴ |
| ۱۳ | ۳۳ | ۰/۹۸ | ۰ | ۴۸۲ |
| ۱۴ | ۳۵ | ۰/۹۸ | ۰ | ۵۶۴ |
| ۱۵ | ۴۴ | ۰/۹۷ | ۰ | ۱۴۲۴ |
| ۱۶ | ۴۸ | ۰/۹۸ | ۰ | ۱۶۱۵ |
| ۱۷ | ۴۹ | ۰/۹۷۹ | ۰ | ۷۰۲ |
| ۱۸ | ۵۲ | ۰/۹۷۳ | ۰ | ۸۵۲ |

| | | | | |
|---|---|---|---|---|
| ۱۹ | ۵۳ | ۰/۹۷ | ۰ | ۶۵۴ |
| ۲۰ | ۵۴ | ۰/۹۸ | ۰ | ۵۹۳ |
| ۲۱ | ۷۳ | ۰/۹۶۷ | ۰ | ۶۸۸ |
| ۲۲ | ۷۸ | ۰/۹۸۱ | ۰ | ۷۵۹ |
| ۲۳ | ۸۹ | ۰/۹۶۶ | ۰ | ۱۵۴۲ |
| ۲۴ | ۹۲ | ۰/۹۷۹ | ۰ | ۵۷۷ |
| ۲۵ | ۹۶ | ۰/۹۷۹ | ۰ | ۵۷۴ |
| ۲۶ | ۱۰۰ | ۰/۹۷۶ | ۰ | ۸۸۸ |
| ۲۷ | ۱۰۸ | ۰/۹۶۹ | ۰ | ۱۸۳۴ |
| ۲۸ | ۱۱۰ | ۰/۹۶۷ | ۰ | ۱۱۹۴ |
| ۲۹ | ۱۲۸ | ۰/۹۸۴ | ۰ | ۱۲۵۶ |
| ۳۰ | ۱۳۳ | ۰/۹۸ | ۰ | ۸۶۱ |
| ۳۱ | ۱۵۰ | ۰/۹۶۵ | ۰ | ۱۵۷۲ |
| ۳۲ | ۱۵۴ | ۰/۹۸۷ | ۰ | ۱۲۲۹ |

**فرضیه شماره ۲.** هرچه میزان آنتروپی متن بیشتر باشد میزان حضور اطلاعات در متن کمتر است.

همان‌طور که پیش از این ذکر شد، آنتروپی کمیتی قابل محاسبه برای هر واحد زبانی است. آنتروپی متن برابر با مجموع آنتروپی واژگان آن متن است. به این ترتیب برای هر مقاله عددی مثبت نشانگر آنتروپی آن مقاله به دست آوردیم. به منظور بررسی میزان حضور اطلاعات در متن کلمات پربار و کم‌بار اطلاعاتی مورد توجه قرار گرفتند. شناسایی این کلمات به این صورت تحقق یافت که ابتدا واژگان هر مقاله به ترتیب آنتروپی ( از بزرگ‌ترین به کوچک‌ترین) مرتب شدند، سپس با توجه به میانگین آنتروپی در هر مقاله دو طیف واژه شناسایی گردید. واژگان کم‌بار همان واژگانی بودند که به واسطه عدد آنتروپی بزرگ در ردیف‌های بالای جدول قرار گرفتند. این واژگان بزرگ‌ترین فراوانی را نیز دارا بودند. پس از آن واژگانی که با آنتروپی و فراوانی کم، در زیر میانگین قرار گرفتند، به عنوان واژگان پربار در نظر گرفته شدند. جدول شماره ۲ مقایسه آنتروپی مقاله و تعداد واژگان پر بار و کم بار اطلاعاتی را نشان می‌دهد:

**جدول شماره ۲. میزان حضور اطلاعات در متن**

| ردیف | کد | میانگین آنتروپی | آنتروپی متن | تعداد واژگان | بالای میانگین | پایین میانگین | نسبت فراوانی واژگان کم‌بار | نسبت فراوانی واژگان پربار | نسبت کم-بار به پربار |
|---|---|---|---|---|---|---|---|---|---|
| ۱ | ۳ | ۰/۰۰۸۶ | ۰/۸۳۲۸ | ۴۹۵۱ | ۳۹۲۵ | ۱۰۲۶ | ۰/۷۹۲۷۶۹ | ۰/۲۰۷۲۳۱ | ۳/۸۲۵۵۳۳ |
| ۲ | ۴ | ۰/۰۰۷ | ۸/۳۴۴۷۹ | ۴۱۸۲ | ۳۰۴۹ | ۱۱۳۶ | ۰/۷۲۹۰۷۷ | ۰/۲۷۱۶۴ | ۲/۶۸۳۹۸۲ |
| ۳ | ۶ | ۰/۰۰۹ | ۷/۸۲۸۲۱۶ | ۴۵۷۴ | ۳۴۵۷ | ۱۱۱۷ | ۰/۷۵۵۷۹۴ | ۰/۲۴۴۲۰۶ | ۳/۰۹۴۹۰۳ |
| ۴ | ۷ | ۰/۰۱۲ | ۸/۱۱۷ | ۳۳۲۹ | ۲۳۸۴ | ۹۴۵ | ۰/۷۱۶۱۳۱ | ۰/۲۸۳۸۶۹ | ۲/۵۲۲۷۵۲ |
| ۵ | ۱۰ | ۰/۰۱۱۷ | ۸/۰۸۴۷ | ۲۰۳۹ | ۱۳۴۴ | ۶۹۵ | ۰/۶۵۹۱۴۷ | ۰/۳۴۰۸۵۳ | ۱/۹۳۳۸۱۶ |
| ۶ | ۱۷ | ۰/۰۱۵۳ | ۷/۷۶۵۴۷ | ۱۸۳۰ | ۱۲۶۵ | ۵۶۵ | ۰/۶۹۱۲۵۷ | ۰/۳۰۸۷۴۳ | ۲/۲۳۸۹۴ |
| ۷ | ۲۰ | ۰/۰۱۳۴ | ۷/۸۵۴۸۵ | ۲۴۵۳ | ۱۷۹۰ | ۶۶۳ | ۰/۷۲۹۷۱۹ | ۰/۲۷۰۲۸۱ | ۲/۶۹۹۸۵۳ |
| ۸ | ۲۱ | ۰/۰۰۵۵ | ۸/۵۵۶۲۱ | ۶۷۴۱ | ۵۰۳۹ | ۱۷۰۲ | ۰/۷۴۷۵۱۵ | ۰/۲۵۲۴۸۵ | ۲/۹۶۰۶۳۱ |
| ۹ | ۲۳ | ۰/۰۱۳۱ | ۷/۱۸۵۲۲ | ۲۸۵۲ | ۲۲۰۶ | ۶۴۶ | ۰/۷۷۳۴۹۲ | ۰/۲۲۶۵۰۸ | ۳/۴۱۴۸۵۵ |
| ۱۰ | ۲۷ | ۰/۰۱۲۹ | ۷/۸۲۷۰۷ | ۲۲۲۴ | ۱۵۰۶ | ۷۱۸ | ۰/۶۷۷۱۵۸ | ۰/۳۲۲۸۴۲ | ۲/۰۹۷۴۹ |
| ۱۱ | ۲۸ | ۰/۰۲۷۶ | ۷/۹۸۷۱ | ۲۲۹۰ | ۱۶۷۳ | ۶۱۷ | ۰/۷۳۰۵۶۸ | ۰/۲۶۹۴۳۲ | ۲/۷۱۱۵۱۲ |
| ۱۲ | ۳۰ | ۰/۰۰۷۱ | ۸/۶۹۴۲۳ | ۳۷۸۵ | ۲۶۱۰ | ۱۱۷۵ | ۰/۶۸۹۵۶۴ | ۰/۳۱۰۴۳۶ | ۲/۲۲۱۲۷۶ |
| ۱۳ | ۳۳ | ۰/۰۱۵۶ | ۷/۵۳۳۴۲ | ۱۸۵۲ | ۱۳۲۵ | ۵۲۷ | ۰/۷۱۵۴۴۳ | ۰/۲۸۴۵۵۷ | ۲/۵۱۴۲۳۴ |
| ۱۴ | ۳۵ | ۰/۰۱۴ | ۷/۹۱۲۹۸ | ۱۷۶۲ | ۱۲۶۱ | ۵۰۱ | ۰/۷۱۵۶۶۴ | ۰/۲۸۴۳۳۶ | ۲/۵۱۶۹۶۶ |
| ۱۵ | ۴۴ | ۰/۰۰۶ | ۸/۸۶۰۳۵ | ۴۷۲۶ | ۳۳۰۹ | ۱۴۱۷ | ۰/۷۰۰۱۶۹ | ۰/۲۹۹۸۳۱ | ۲/۳۳۵۲۱۲ |
| ۱۶ | ۴۸ | ۰/۰۰۵۳ | ۸/۸۶۰۰۲ | ۵۵۸۳ | ۴۰۹۳ | ۱۴۹۰ | ۰/۷۳۳۱۱۸ | ۰/۲۶۶۸۸۲ | ۲/۷۴۶۹۷۴ |
| ۱۷ | ۴۹ | ۰/۰۱۱۶ | ۸/۱۵۹۹۹ | ۲۲۵۳ | ۱۶۲۹ | ۶۲۴ | ۰/۷۲۳۰۳۶ | ۰/۲۷۶۹۶۴ | ۲/۶۱۰۵۷۸ |
| ۱۸ | ۵۲ | ۰/۰۰۹۵ | ۸/۰۹۳۴۴ | ۳۷۱۹ | ۲۸۳۲ | ۸۸۷ | ۰/۷۶۱۴۹۵ | ۰/۲۳۸۵۰۵ | ۳/۱۹۲۷۸۴ |
| ۱۹ | ۵۳ | ۰/۰۱۱۸ | ۸/۰۴۷۰۸ | ۲۷۸۲ | ۲۰۶۶ | ۷۱۶ | ۰/۷۴۲۶۳۱ | ۰/۲۵۷۳۶۹ | ۲/۸۸۵۴۷۲ |
| ۲۰ | ۵۴ | ۰/۰۱۳۶ | ۷/۵۹۸۸۲ | ۱۷۴۴ | ۱۲۱۱ | ۵۳۳ | ۰/۶۹۴۳۸۱ | ۰/۳۰۵۶۱۹ | ۲/۲۷۲۰۴۸ |
| ۲۱ | ۷۳ | ۰/۰۱۱۴ | ۷/۸۶۹۶۸ | ۲۶۵۶ | ۱۹۰۹ | ۷۴۷ | ۰/۷۱۸۷۵ | ۰/۲۸۱۲۵ | ۲/۵۵۵۵۵۶ |
| ۲۲ | ۷۸ | ۰/۰۱۰۲ | ۸/۳۱۷۱۱ | ۳۵۴۴ | ۲۷۴۱ | ۸۰۳ | ۰/۷۷۳۴۲ | ۰/۲۲۶۵۸ | ۳/۴۱۳۴۵۲ |
| ۲۳ | ۸۹ | ۰/۰۰۵۶ | ۸/۶۴۷۹۱ | ۵۰۶۱ | ۳۶۲۶ | ۱۴۳۵ | ۰/۷۱۶۴۵۹ | ۰/۲۸۳۵۴۱ | ۲/۵۲۶۸۲۷ |
| ۲۴ | ۹۲ | ۰/۰۱۳۸ | ۷/۹۵۱۷۶ | ۲۰۹۲ | ۱۴۵۷ | ۶۳۵ | ۰/۶۹۶۴۶۳ | ۰/۳۰۳۵۳۷ | ۲/۲۹۴۴۹۱ |
| ۲۵ | ۹۶ | ۰/۰۱۳۹ | ۷/۹۷۳۷۷ | ۱۶۶۰ | ۱۱۳۵ | ۵۲۵ | ۰/۶۸۳۷۳۵ | ۰/۳۱۶۲۶۵ | ۲/۱۶۱۹۰۵ |
| ۲۶ | ۱۰۰ | ۰/۰۰۹۵ | ۸/۴۲۸۴۵ | ۲۷۷۶ | ۱۹۴۳ | ۸۳۳ | ۰/۶۹۹۹۲۸ | ۰/۳۰۰۰۷۲ | ۲/۳۳۲۵۳۴ |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ۲۷ | ۱۰۸ | ۰/۰۰۴۸ | ۸/۷۴۰۱۵ | ۸۱۱۶ | ۶۰۶۵ | ۲۰۵۱ | ۰/۷۴۷۲۸۹ | ۰/۲۵۲۷۱۱ | ۲/۹۵۷۰۸۹ |
| ۲۸ | ۱۱۰ | ۰/۰۰۷۱ | ۸/۴۱۱۰۴ | ۴۴۳۲ | ۳۱۰۱ | ۱۳۳۱ | ۰/۶۹۹۶۸۴ | ۰/۳۰۰۳۱۶ | ۲/۳۲۹۸۲۶ |
| ۲۹ | ۱۲۸ | ۰/۰۰۶۹ | ۸/۶۵۲۱۹ | ۳۵۵۲ | ۲۳۲۶ | ۱۲۲۶ | ۰/۶۵۴۸۴۲ | ۰/۳۴۵۱۵۸ | ۱/۸۹۷۲۲۴ |
| ۳۰ | ۱۳۳ | ۰/۰۰۹۸ | ۸/۴۰۳۹۸ | ۲۷۴۹ | ۱۹۶۲ | ۷۸۷ | ۰/۷۱۳۷۱۴ | ۰/۲۸۶۲۸۶ | ۲/۴۹۳۰۱ |
| ۳۱ | ۱۵۰ | ۰/۰۰۵۷ | ۸/۸۸۷۸ | ۵۰۴۴ | ۳۵۴۸ | ۱۴۹۶ | ۰/۷۰۳۴۱ | ۰/۲۹۶۵۹ | ۲/۳۷۱۶۵۸ |
| ۳۲ | ۱۵۴ | ۰/۰۰۷ | ۸/۵۴۸۸۵ | ۳۷۶۳ | ۲۶۰۰ | ۱۱۶۳ | ۰/۶۹۰۹۳۸ | ۰/۳۰۹۰۶۲ | ۲/۲۳۵۵۹۷ |

به منظور آزمون فرض فوق ابتدا نسبت واژگانی که آنتروپی آن‌ها بالای میانگین بود به تعداد کل واژگان برای هر مقاله محاسبه شد، این محاسبه برای تعداد واژگانی که آنتروپی آن‌ها پایین میانگین بود نیز حساب شد. سپس بین این دو تعداد آزمون همبستگی انجام شد.

**جدول شماره ۳. اختلاف معناداری تعداد واژگان بالا و پایین میانگین**

| ضریب همبستگی | Sig. (2-tailed) | N |
|---|---|---|
| ۱- | ۰ | ۳۲ |

نتایج آزمون نشان داد که بین تعداد واژگان بالای میانگین (نمایانگر واژگان کم بار اطلاعاتی) و واژگان پایین میانگین (نمایانگر واژگان پربار اطلاعاتی) در همه مقالات رابطه معنادار معکوسی وجود دارد. مرحله دوم آزمون فرض، بررسی رابطه این دو سطح از متغیر میزان اطلاعات متن با متغیر آنتروپی متن بود. به این منظور از آزمون رگرسیون چند متغیری[1] استفاده شد. جدول شماره ۴ حاصل انجام این آزمون در نرم‌افزار اس.پی.اس.اس[2] است.

**جدول شماره ۴. رابطه اطلاعات متن با آنتروپی**

| Model | ضرایب غیراستاندارد | | ضرایب استاندارد | T | Sig. |
|---|---|---|---|---|---|
| | B | Std. Error | Beta | | |
| (Constant) | ۸/۲۴۲ | ۰/۹۴۳ | | ۸/۷۳۷ | ۰ |
| بالای میانگین | ۷/۴۵۶ | ۰/۴۰۶ | ۱/۲۵ | ۷/۷۵۷ | ۰ |
| پایین میانگین | ۰ | ۰/۰۰۱ | ۰/۰۹۲- | ۰/۴۴۳- | ۰/۶۶۱ |

ارزشP به دست آمده در دو سطح واژگان بالا و پایین میانگین آنتروپی، نشان‌دهنده ارتباط معنی‌دار بین اطلاعات متن و آنتروپی آن در سطح واژگان بالای میانگین است. این ارتباط در سطح واژگان پایین میانگین معنادار نیست. درواقع آنتروپی متن با تعداد واژگان کم‌بار اطلاعاتی رابطه دارد و می‌توان گفت هرچه تعداد این واژگان بیشتر باشد آنتروپی متن نیز بالاتر است.

1. Multiple Regression
2. SPSS

**فرضیه شماره ۳.** هرچه تعداد حالات ممکن یک واژه کمتر باشد بار اطلاعاتی آن واژه بیشتر است.

تعداد حالات ممکن یک واژه، شکل‌های مختلف ظهور یک واژه در متن اعم از حالات اسمی، فعلی، قیدی و ... است. به طور مثال حالات ممکن واژه *تدارک*، در یک مقاله شامل واژه‌های *تدارک*، *تدارک دیدن*، *تدارکات*، *تدارک دید*، خواهد بود. پس از شمارش حالات ممکن هر واژه یک ستون به آن اختصاص یافت و سپس فراوانی کلیه حالات ممکن یک واژه تجمیع گردید. به این ترتیب واژه‌ای با ۶ حالت ممکن تنها به یک شکل در جدول ظاهر می‌شود، در حالی که فراوانی سایر حالات به فراوانی آن افزوده گشته است. در مرحله بعد واژگان غیرمجاز از جداول حذف شد. چشم‌پوشی از واژگان غیرمجاز به این دلیل انجام شد که این واژگان بیشترین فراوانی را دارا بوده در عین حال حالات ممکن متعددی ندارند و این امر می‌توانست صحت آزمون همبستگی را تحت الشعاع قرار دهد. پس از این مرحله به بررسی رابطه این ستون با ستون احتمال رخداد پرداختیم.

**جدول شماره ۵. بررسی ضریب همبستگی آنتروپی واژه با حالات ممکن آن**

| ردیف | کد مقاله | ضریب همبستگی | Sig. (2-tailed) | تعداد واژه |
|---|---|---|---|---|
| ۱ | ۳ | ۰/۵۲۴ | ۰ | ۹۳۹ |
| ۲ | ۴ | ۰/۴۲۴ | ۰ | ۱۱۹۲ |
| ۳ | ۶ | ۰/۴۷۸ | ۰ | ۸۷۳ |
| ۴ | ۷ | ۰/۴۸۶ | ۰ | ۸۵۵ |
| ۵ | ۱۰ | ۰/۵۰۹ | ۰ | ۷۰۰ |
| ۶ | ۱۷ | ۰/۵۰۲ | ۰ | ۵۰۰ |
| ۷ | ۲۰ | ۰/۵۵۸ | ۰ | ۵۸۵ |
| ۸ | ۲۱ | ۰/۵۶۲ | ۰ | ۱۵۶۶ |
| ۹ | ۲۳ | ۰/۳۳۸ | ۰ | ۵۴۹ |
| ۱۰ | ۲۷ | ۰/۴۷ | ۰ | ۶۰۹ |
| ۱۱ | ۲۸ | ۰/۴۹۳ | ۰ | ۶۸۶ |
| ۱۲ | ۳۰ | ۰/۵۷۷ | ۰ | ۱۲۲۴ |
| ۱۳ | ۳۳ | ۰/۴۶۱ | ۰ | ۴۸۲ |
| ۱۴ | ۳۵ | ۰/۶۰۳ | ۰ | ۵۶۴ |
| ۱۵ | ۴۴ | ۱ | ۰ | ۱۴۲۴ |
| ۱۶ | ۴۸ | ۰/۳۴۲ | ۰ | ۱۶۱۵ |
| ۱۷ | ۴۹ | ۰/۶۶۸ | ۰ | ۷۰۲ |

| ردیف | کد مقاله | ضریب همبستگی | Sig. (2-tailed) | تعداد واژه |
|---|---|---|---|---|
| ۱۸ | ۵۲ | ۰/۴۴ | ۰ | ۸۵۲ |
| ۱۹ | ۵۳ | ۰/۳۸۹ | ۰ | ۶۵۴ |
| ۲۰ | ۵۴ | ۰/۵۴۹ | ۰ | ۵۹۳ |
| ۲۱ | ۷۳ | ۰/۴۶۱ | ۰ | ۶۸۸ |
| ۲۲ | ۷۸ | ۰/۳۸۷ | ۰ | ۷۵۹ |
| ۲۳ | ۸۹ | ۰/۵۵۵ | ۰ | ۱۵۴۲ |
| ۲۴ | ۹۲ | ۰/۵۹ | ۰ | ۵۷۷ |
| ۲۵ | ۹۶ | ۰/۵۸۶ | ۰ | ۵۷۴ |
| ۲۶ | ۱۰۰ | ۰/۵۲۲ | ۰ | ۸۸۸ |
| ۲۷ | ۱۰۸ | ۰/۴۵۵ | ۰ | ۱۸۳۴ |
| ۲۸ | ۱۱۰ | ۰/۶۰۲ | ۰ | ۱۱۹۴ |
| ۲۹ | ۱۲۸ | ۰/۵۱۸ | ۰ | ۱۲۵۶ |
| ۳۰ | ۱۳۳ | ۰/۶۰۳ | ۰ | ۸۶۱ |
| ۳۱ | ۱۵۰ | ۰/۴۹۵ | ۰ | ۱۵۷۲ |
| ۳۲ | ۱۵۴ | ۰/۴۴۷ | ۰ | ۱۲۲۹ |

با توجه به مقدار P به دست آمده از تک تک مقالات فرض پژوهش پذیرفته می‌شود:

$$\text{Sig.(2-tailed)} = \text{P-Value} < 0.05$$

بررسی جدول شماره ۵نشان‌دهنده پذیرش فرض آزمون است. طبق یافته‌هایاین جدول ارتباط معنی‌داری میان تعداد حالات ممکن و آنتروپی وجود دارد. ضریب همبستگی پیرسون ۰/۳ تا ۱ می‌باشد و این مطلب نشان‌دهنده همبستگی متوسط تا قوی است.

**فرضیه شماره ۴.** بین طول کلمه و بار اطلاعاتی آن رابطه وجود دارد.

هدف از طرح این فرضیه بررسی این موضوع است که آیا واژگان با طول‌های متفاوت، بار اطلاعاتی یکسانی دارند یا نه؟ به عبارت دیگر آیا تک تک واژگان صرف‌نظر از متنی که در آن ظاهر می‌شوند به میزان یکسانی اطلاعات منتقل می‌نمایند و آنتروپی مشابهی دارند. این فرضیه نیز باید در سطح تک تک واژه‌های مقالات نمونه بررسیمی‌شد. متغیرهای مورد بررسی در این فرضیه طول کلمه و آنتروپی آن می‌باشد. طول کلمه همان تعداد کاراکترهای یک واژه است که پس از انجام عملیات ویرایش و آماده‌سازی متون به صورت جداول با استفاده از قابلیت Word Count محاسبه و در ستونی جداگانه در صفحه اکسل به نمایش درآمد. آنتروپی کلیه واژگان نیز همان‌گونه که قبلاً توضیح داده شد محاسبه

گردید.

**جدول شماره ۶. رابطه آنتروپی واژه با طول آن**

| ردیف | کد مقاله | ضریب همبستگی | Sig. (2-tailed) | تعداد واژه |
|---|---|---|---|---|
| ۱ | ۳ | −۰/۲۳۹ | ۰ | ۹۳۹ |
| ۲ | ۴ | −۰/۱۹۷ | ۰ | ۱۱۹۲ |
| ۳ | ۶ | −۰/۲۱۶ | ۰ | ۸۳۷ |
| ۴ | ۷ | −۰/۲۱۱ | ۰ | ۸۵۷ |
| ۵ | ۱۰ | −۰/۲۳۳ | ۰ | ۶۹۹ |
| ۶ | ۱۷ | −۰/۲۱۱ | ۰ | ۵۰۰ |
| ۷ | ۲۰ | −۰/۱۹۷ | ۰ | ۵۸۶ |
| ۸ | ۲۱ | −۰/۲۱۳ | ۰ | ۱۵۶۶ |
| ۹ | ۲۳ | −۰/۰۹۴ | ۰/۰۲۸ | ۵۴۹ |
| ۱۰ | ۲۷ | −۰/۲۱۹ | ۰ | ۶۰۹ |
| ۱۱ | ۲۸ | −۰/۲۰۲ | ۰ | ۶۸۶ |
| ۱۲ | ۳۰ | −۰/۱۲۶ | ۰ | ۱۲۲۳ |
| ۱۳ | ۳۳ | −۰/۲۵۲ | ۰ | ۴۸۲ |
| ۱۴ | ۳۵ | −۰/۲۱۲ | ۰ | ۵۶۳ |
| ۱۵ | ۴۴ | −۰/۱۸۶ | ۰ | ۱۴۸۱ |
| ۱۶ | ۴۸ | −۰/۱۸۳ | ۰ | ۱۶۱۵ |
| ۱۷ | ۴۹ | −۰/۲۲۶ | ۰ | ۷۰۲ |
| ۱۸ | ۵۲ | −۰/۲۲۶ | ۰ | ۸۵۲ |
| ۱۹ | ۵۳ | −۰/۲۳۴ | ۰ | ۶۴۵ |
| ۲۰ | ۵۴ | −۰/۲۲۳ | ۰ | ۵۹۲ |
| ۲۱ | ۷۳ | −۰/۱۸۶ | ۰ | ۶۸۸ |
| ۲۲ | ۷۸ | −۰/۱۰۶ | ۰/۰۰۳ | ۷۵۹ |
| ۲۳ | ۸۹ | −۰/۱۸ | ۰ | ۱۵۴۲ |
| ۲۴ | ۹۲ | −۰/۲۲۲ | ۰ | ۵۷۷ |
| ۲۵ | ۹۶ | −۰/۲۱۲ | ۰ | ۵۷۴ |
| ۲۶ | ۱۰۰ | −۰/۱۸۳ | ۰ | ۸۸۸ |
| ۲۷ | ۱۰۸ | −۰/۱۸۸ | ۰ | ۱۸۳۴ |
| ۲۸ | ۱۱۰ | −۰/۱۷۹ | ۰ | ۱۱۹۴ |

| | | | | |
|---|---|---|---|---|
| ۲۹ | ۱۲۸ | −۰/۲ | ۰ | ۱۲۵۶ |
| ۳۰ | ۱۳۳ | ۰/۱۹۷ | ۰ | ۸۶۱ |
| ۳۱ | ۱۵۰ | ۰/۱۹۳ | ۰ | ۱۵۷۲ |
| ۳۲ | ۱۵۴ | ۰/۲۱۹ | ۰ | ۱۲۲۹ |

با توجه به جدول شماره ۶ در تمامی مقالات موردبررسی ارزش P (Sig.) کوچک‌تر از ۰/۰۵ است که این مطلب بیانگر وجود رابطه میان طول کلمه و بار اطلاعاتی آن می‌باشد. ضریب همبستگی به دست آمده نیز در هر ۳۲ نمونه مورد بررسی عددی منفی است که رابطه‌ای معکوس بین آنتروپی واژه و طول آن را نشان می‌دهد. به این ترتیب، هرچه واژه‌ای کوتاه‌تر باشد آنتروپی واژه بیشتر و بنابراین اطلاعات واژه کمتر خواهد بود.

**فرضیه شماره ۵.** مقدار بار اطلاعاتی متون در حوزه‌های مختلف علمی متفاوت می‌باشد.

هدف از مطرح نمودن فرضیه فوق، بررسی این موضوع است که آیا بافت‌های واژگانی با درون‌مایه متفاوت، میزان اطلاعات یکسانی منتقل می‌سازند یا برخی متون با موضوعی خاص نسبت به متون دیگر با موضوعی متفاوت اطلاعات بیشتری دارا هستند؟همان‌طور که پیش از این گفته شد کمیت آنتروپی قابل محاسبه برای کلیه واحدهای زبانی است؛ پس از محاسبه آنتروپی تک تک واژگان یک مقاله با استفاده از فرمول می‌توان آنتروپی متن را قابل محاسبه ساخت:

$$H = -\sum P(i)LogP(i)$$

علامت سیگما (Σ) در فرمول نشان‌دهنده این است که آنتروپی متن حاصل جمع آنتروپی کلیه واژگان آن متن است. به این ترتیب برای هر مقاله یک عدد مثبت نشانگر آنتروپی متن به دست می‌آید. بدیهی است که متغیر مورد بررسی در این فرضیه آنتروپی $H$ هر مقاله است. سطح مورد آزمون نیز چهار حوزه است که میانگین آنتروپی هر حوزه در جدول شماره ۷ ارائه شده است. قابل ذکر است که سطح مورد توجه در این فرضیه حوزه است، به این ترتیب ۳۲۰ مقاله اولیه مورد بررسی قرار گرفتند.

**جدول شماره ۷. میانگین آنتروپی در حوزه‌های علمی**

| ردیف | حوزه علمی | تعداد مقالات | میانگین آنتروپی |
|---|---|---|---|
| ۱ | ادبیات و علوم انسانی | ۱۵۴ | ۸/۴۵ |
| ۲ | علوم پایه | ۳۰ | ۷/۹۸۹ |
| ۳ | فنی و مهندسی | ۳۵ | ۷/۵۹۰۶۴ |
| ۴ | کشاورزی | ۱۰۱ | ۷/۹۲۳ |

به منظور بررسی این فرضیه از آزمون تحلیل واریانس[1] یک طرفه استفاده گردید. از این آزمون در مواقعی استفاده می‌شود که آمارگر قصد دارد میانگین‌های سه جامعه یا بیشتر را با یکدیگر مقایسه کند یا به برآورد و مقایسه میانگین یک صفت در چند جامعه و همچنین وجود یا عدم وجود تفاوت معنی‌دار در بین نمونه‌های یک جامعه بپردازد ( هویدا، ۱۳۷۸). با توجه به اینکه هدف از این فرضیه بررسی آنتروپی در چهار حوزه ادبیات و علوم انسانی، فنی و مهندسی، علوم پایه و کشاورزی می‌باشد بنابراین فرضیه با آزمون تحلیل واریانس بررسی شد.

**جدول شماره ۸. آزمون تحلیل واریانس میانگین آنتروپی حوزه‌های علمی**

| ANOVA | مجموع مربعات خطا | *Df* | *F* | *Sig.* |
|---|---|---|---|---|
| بین گروهی | ۳۰/۶۹۳ | ۳ | ۱۰/۸۸ | ۰ |
| درون گروهی | ۲۹۷/۱۴۶ | ۳۱۶ | | |

با توجه به جدول حاصل از آزمون ANOVA، P-Value از ۰/۰۵ کوچک‌تر است و فرض پژوهش پذیرفته می‌شود (Sig.= P-Value=0.000<0.05)؛ همچنین نسبت F در صورتی که مساوی یا کمتر از یک باشد نتیجه معنی‌دار نیست و بالعکس هرچه بزرگ‌تر باشد اثر متغیر مستقل بر داده‌ها بیشتر است. البته میزان بزرگی نسبت F به حدی که بتوانیم آن را معنادار بدانیم بستگی به ارزش P دارد و باید کمتر از ۰/۰۵ باشد تا بتوان F را معنادار دانست (بریس، کمپ و سنلگار، ۱۳۹۱). با این شرح نسبت F بین گروه‌ها نشانگر معناداری اختلاف میان آنتروپی حوزه‌هاست. به منظور بررسی رابطه بین حوزه‌ها به صورت دو به دو از آزمون تعقیبی LSD استفاده شد. جدول شماره ۹ نشان‌دهنده میزان اختلاف در اطلاعات چهار حوزه علمی است.

**جدول شماره ۹. اختلاف آنتروپی در حوزه‌های علمی**

| حوزه | | اختلاف میانگین | انحراف استاندارد | Sig. |
|---|---|---|---|---|
| ادبیات و علوم انسانی | علوم پایه | ۰/۴۶۰۷۴ | ۰/۱۹۳۵۲ | ۰/۰۱۸ |
| | مهندسی | ۰/۸۵۸۴۳ | ۰/۱۸۱۵۸ | ۰ |
| | کشاورزی | ۰/۵۲۵۷۳ | ۰/۱۲۴۱۶ | ۰ |
| علوم پایه | ادبیات و علوم انسانی | ۰/۴۶۰۷۴ | ۰/۱۹۳۲۵ | ۰/۰۱۸ |
| | مهندسی | ۰/۳۹۷۶۹ | ۰/۲۴۱۲۷ | ۰/۱ |
| | کشاورزی | ۰/۰۶۴۹۹ | ۰/۲۰۱۶۳ | ۰/۷۴۷ |
| مهندسی | ادبیات و علوم انسانی | ۰/۸۵۸۴۳ | ۰/۱۸۱۵۸ | ۰ |

1. Analysis of Variance (ANOVA)

| | | | | |
|---|---|---|---|---|
| | علوم پایه | ۰/۳۹۷۶۹ | ۰/۲۴۱۲۷ | ۰/۱ |
| | کشاورزی | ۰/۳۳۲۷ | ۰/۱۹۰۲ | ۰/۰۸۱ |
| کشاورزی | ادبیات و علوم انسانی | ۰/۵۲۵۷۳ | ۰/۱۲۴۱۶ | ۰ |
| | علوم پایه | ۰/۰۶۴۹۹ | ۰/۲۰۱۶۳ | ۰/۷۴۷ |
| | مهندسی | ۰/۳۳۲۷ | ۰/۱۹۰۲ | ۰/۰۸۱ |

با بررسی ارزش P حاصل مقایسه میانگین حوزه‌ها فرض پژوهش پذیرفته می‌شود. در نهایت می‌توان گفت که حوزه ادبیات و علوم انسانی با سه حوزه دیگر از نظر اطلاعاتی که منتقل می‌نماید متفاوت است.

## نتیجه

بررسی آنتروپی و بار اطلاعاتی واژه در متون فارسی از چند جنبه در پژوهش حاضر مورد توجه قرار گرفت. اولین ویژگی مورد مطالعه یک واژه در متن احتمال رخداد واژه و رابطه آن با میزان اطلاعات واژه بود. شانون به صراحت در نظریه معروف خود تعریف کرد که اطلاعات I در یک پیام به طور معکوس با احتمال آن رابطه معکوس دارد (Bartlett, 2007). نتایج حاصل از پژوهش‌های مختلف نشان‌دهنده رابطه معکوس باراطلاعاتی یک واژه با احتمال رخداد آن است. مک‌کی[1] (۲۰۰۳) محتوای اطلاعاتی ۲۷ رخداد ممکن را هنگامی‌که یک کاراکتر به طور تصادفی از متون انگلیسی انتخاب می‌شود، مورد محاسبه قرار داد. به این ترتیب محتوای اطلاعاتی رخداد حرف z برابر ۱۰.۴ و حرف e برابر ۳.۵ بیت است. در این رابطه احتمال رخداد با میزان اطلاعات شانون رابطه معکوس دارد. بدیهی است هرچه احتمال وقوع یک پیشامد بیشتر باشد آن پیشامد اطلاعات کمتری را انتقال می‌دهد و این نتیجه در آزمون رابطه بین دو متغیر آنتروپی و احتمال رخداد در پژوهش حاضر نیز به دست آمد، به این معنا که واژه‌هایی که با آنتروپی بالایی در متن ظاهر شدند بیشترین احتمال رخداد را داشتند. فیریسون، راینر و پیکرینگ[2] (۲۰۰۵) در پژوهش خود با بررسی هم وقوعی واژگان ( احتمال رخداد دو واژه با هم)، معتقدند یک واژه با قابلیت پیش‌بینی‌پذیری بالاواژه‌ای است که با احتمال رخداد بالایی در متن ظاهر می‌شود. همان‌طور که گفته شد طبق اصول شانون واژه‌ای که پیش‌بینی‌پذیری زیادی دارد درحالی‌که آنتروپی بالایی دارد اطلاعات اندکی خواهد داشت. بنابراین طبق رابطه سه مفهوم مذکور خواهیم داشت:

1. MacKay
2. Frisson, Rayner & Pickering

پیش‌بینی‌پذیری زیاد ↔ احتمال رخداد بالا ↔ افزایش آنتروپی ↔ اطلاعات اندک

بنابر آنچه گفته شد میزان اطلاعات واژه با احتمال رخداد آن رابطه‌ای معکوس دارد.

در فرضیه دوم پژوهش حاضر، به میزان آنتروپی متن و رابطه آن با حضور اطلاعات در همان متن پرداختیم. همان‌گونه که در رابطه با آزمون فرض دوم پژوهش گفته شد، آنتروپی یک مجموعه حاصل مجموع آنتروپی تک‌تک اجزای آن مجموعه است. درواقع اگر آنتروپی معادل عدم وجود اطلاعات در یک سامانه باشد، بنابراین آنتروپی بالا نشان‌دهنده اطلاعات اندک آن خواهد بود. آزمون فرض بین متغیر آنتروپی متن و اطلاعات آن نشان دادکه آنتروپی متن با تعداد واژگان بالای میانگین (واژگان کم‌بار) رابطه دارد؛درحالی‌که با تعداد واژگان پایین میانگین (واژگان پربار) این رابطه تأیید نشد. البته شناسایی واژه‌های پربار و کم بار اطلاعاتی متن نیازمند تعریف شاخص دقیق‌تر دیگری است؛ ولی با توجه به یافته‌هامی‌توان گفت آنتروپی متن تنها با واژگانی که به طور قطعی واژگان کم‌بار مقاله هستند رابطه داشت.

بار اطلاعاتی واژه از جنبه تعداد حالات ممکن آن در متن نیز قابل بررسی است. هدف از بررسی رابطه تعداد حالات ممکن واژه و بار اطلاعاتی آن، بررسی این موضوع است که آیا واژه‌ای که به شکل‌های مختلف (واژه‌های هم‌خانواده و هم‌ریشه) در یک متن ظاهر می‌شود میزان اطلاعات بیشتری منتقل می‌نماید یا خیر. یو، وو، و هاوی[1] (۲۰۰۸) در فرآیند ابهام‌زدایی معنایی از واژه به رابطه حالات مختلف یک واژه و آنتروپی اشاره کردند. به نظر آنان هرچقدر ابهام معنایی یک واژه بالا باشد و به عبارت بهتر واژه دارای معانی و اشکال متعددی باشد آنتروپی بالاتری برای آن واژه خواهیم داشت. همچنین بار اطلاعاتی آن واژه بیشتر است چرا که آن واژه غیرقابل پیش‌بینی‌تر است و محتوای اطلاعاتی بیشتری به همراه دارد.

بررسی رابطه طول واژه و میزان اطلاعات آن نشان داد که هرچه یک واژه کوتاه‌تر باشد میزان اطلاعات کمتری نیز خواهد داشت. به این ترتیب رابطه آنتروپی و طول واژه رابطه‌ای معکوس است. طبق نظریه شانون یک زبان کارآمد میزان اطلاعاتی نزدیک به ظرفیت کانال انتقال می‌دهد، به این ترتیب یک رابطه غیرخطی بین منفی لگاریتم احتمال واژه و طول آن وجود دارد و این رابطه همان فرمول آنتروپی شانون است. کانچو و مارتین[2] (۲۰۱۱) نیز با بررسی متونی که به طور تصادفی و بی‌هدف تایپ می‌شوندبه رابطه خطی بین طول واژه$\ell$ و محتوای اطلاعاتی $I(\ell)$آن اشاره کردند. بنابر آنچه گفته شد نتیجه حاصل حکایت از وجود رابطه بین طول کلمه و بار اطلاعاتی آن داشت.

1. Yu, Wu & Hovy
2. Cancho & Martin

درنهایت، نتایج محاسبه بار اطلاعاتی متون حوزه‌های علمی مختلف در چهار حوزه علمی ادبیات و علوم انسانی، علوم پایه، فنی و مهندسی و کشاورزی نشان داد بیشترین آنتروپی و کمترین اطلاعات در حوزه علوم انسانی نسبت به سایر حوزه‌هامی‌باشد. پس از حوزه علوم انسانی، حوزه علوم پایه، کشاورزی، و فنی و مهندسی به ترتیب از بالاترین آنتروپی تا کمترین میزان آن قرار می‌گیرند. این مسئله به دلیل وجود مقالات طولانی و پرتکرار حوزه علوم انسانی است. آنچه در مقالات این حوزه به چشم می‌خورد فراوانی قابل توجه واژگان، و به دنبال آن احتمال رخداد بالا و بار اطلاعاتی اندك واژه می‌باشد. ضمن اینکه با افزایش طول مقاله احتمال رخداد واژگان غیرمجاز نیز افزایش یافته و این خود عاملی در افزایش میانگین آنتروپی حوزه است. بنابراین می‌توان گفت بار اطلاعاتی واژه وابسته به متن است و مقدار اطلاعات متون در حوزه‌های مختلف علمی متفاوت می‌باشد.

## کتابنامه


بریس، نیکلا، کمپ ریچارد، و سنلگار، رزمری (۱۳۹۱). *تحلیل داده‌های روانشناسی با برنامه اس پی اس اس*. ترجمه خدیجه علی‌آبادی و علی صمدی. تهران: دوران

حری، عباس (۱۳۸۱). *دائره‌المعارف کتابداری و اطلاع‌رسانی*. تهران: مرکز اسناد و کتابخانه ملی ایران

داورپناه، محمدرضا، و بلندیان، صدیقه (۱۳۸۶). تحلیل متن مقالات فارسی و امکان نمایه سازی ماشینی آن‌ها براساس قانون زیف. *فصلنامه پژوهش در مسائل تعلیم و تربیت: ویژه نامه کتابداری و اطلاع رسانی*، دور دوم

سنجی، مجیده، و داورپناه، محمدرضا (۱۳۸۸). شناسایی واژه‌های غیرمفهومی (رایج) در نمایه سازی خودکار مدارك فارسی. *فصلنامه کتابداری و اطلاع رسانی*، ۱۲(۴۸)، ۲۳– ۳۵

میرزایی، اردوان. (۱۳۸۵). حشو در زبان با رویکرد نظریه اطلاعات. *پیک نور*، ۴(۴)، ۴۰–۴۸

وحیدیان کامیار، تقی (۱۳۷۹). *دستور زبان فارسی۱*. تهران: سمت

هاشمی، محسن، و ساوجی، محمدحسن (۱۳۸۶). فشرده سازی متن فارسی با استفاده از الگوریتم‌های حسابی و هافمن و مقایسه آن با فشرده سازی متن انگلیسی. *مجموعه مقالات پانزدهمین کنفرانس مهندسی برق ایران*. تهران

هویدا، علیرضا (۱۳۷۸). *آمار و روش‌های کمی در کتابداری و اطلاع‌رسانی*. تهران: سمت.

Bartlett, M. S. (2007). Information Maximization in Face Processing. *Neurocomputing, 70*(13–15), 2204-2217.

Cancho, R., & Martín, F. M. d. P. (2011).Information content versus word length in random typing. *Journal of Statistical Mechanics: Theory and Experiment, 2011*(12), L12002.

Caraballo, S. A., & Charniak, E. (1999). Determining The Specificity Of Nouns From Text. *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.*

Fragos, K., Maistros, Y., & Skourlas, C. (2005).A weighted maximum entropy language model for text classification. *NLUCS, 55-67.*

Frisson, S., Rayner, K., & Pickering, M. J. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *JExpPsychol Learn MemCogn, 31*(5), 862-877.

Genzel, D., & Charniak, E. (2002).Entropy rate constancy in text. *Paper presented at the Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, Pennsylvania.

Hegazi, N., Ali, N., & Abed, E. (1987). Information content in textual data: Revisited for Arabic text. *Journal of the American Society for Information Science, 38*(2), 133-137.

Kireyev, K. (2009). Semantic-Based Estimation of Term Informativeness. *The 2009 Annual Conference of the North American Chapter of the ACL*. Boulder, Colorado

Lin, L., & Yu-Shu, L. (2005). Research and realization of naive Bayes English text classification method based on base noun phrase identification. *Paper presented at ITI 3rd International Conference*, Cairo.

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1, 309-317.

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*.Cambridge: CambridgeUniversity Press.

Manning, C. D. & Schutze, H. (1999).*Foundations of Statistical Natural Language Processing*, Cambridge:MIT Press.

Melamed, I.D. (1997). Measuring Semantic Entropy. *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics.*

Montemurro, M. A. & Zanette, D. H. (2001).Entropic analysis of the role of words in literary texts. *Advances in Complex Systems.5(1)*, 7-17.

Nemirovsky, D., & Dobrynin, V. (2008).Word importance discrimination using context information. *In Proceedings TREC*

Ryu, P. M., & Choi, K. S. (2004). Determining the Specificity of Terms based on Information Theoretic Measures. *Paper presented at Poster Session of 3rd International Workshop on Computational Terminology*. Daejeon, Korea

Tomokiyo, T., & Hurst, M. (2003). A language model approach to keyphrase extraction. *Paper presented at the Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment* - Volume 18, Sapporo, Japan.

Weber, P. R. (1990). *Basic content analysis*. Newbury Park, Calif.

Yu, L.-C., Wu, C.-H., & Hovy, E. (2008). OntoNotes: corpus cleanup of mistaken agreement using word sense disambiguation. *Paper presented at the Proceedings of the 22nd International Conference on Computational Linguistics* - Volume 1, Manchester, United Kingdom.

Zou F., Wang F.L., Deng X., & Han S. (2006), Automatic Identification of Chinese Stop Words. *A special issue on Advances in Natural Language Processing of the journal Research on Computing Science* (p 151-162).